回归分析阅读笔记

核心概念

- **统计建模工具**：通过建立变量间数学关系预测或解释目标变量的方法
- **变量角色**：
  - 因变量（被解释变量）
  - 自变量（解释变量）
- **核心价值**：揭示变量间数量变化规律，支持预测和决策

主要回归类型

类型	适用场景	输出特征	典型应用
一元线性回归	适用于单个自变量的线性关系模型	一个自变量，常用于简单趋势预测	房价预测、销售预测等
多元线性回归	适用于多个自变量的线性关系模型	多个自变量，适用于更复杂的数据集	经济预测、市场分析等
岭回归	适用于存在多重共线性问题的场景	添加正则化项以减少过拟合	回归分析中的防过拟合应用
Lasso回归	适用于特征选择，减少模型复杂度	L1正则化，自动执行特征选择	高维数据分析、特征选择等
本质线性模型	适用于数据之间存在线性关系的情况	模型本身为线性	预测线性关系的应用
非本质线性模型	适用于数据之间存在线性模型之外的关系	通过非线性变换进行拟合	复杂系统建模、非线性回归等

4.1回归分析概述

回归分析（**Regression Analysis**）是研究一个随机变量（因变量）与其他变量（自变量）之间的关系，常用于分析自变量与因变量之间的统计学关系。其主要目的是建立自变量与因变量之间的数学模型，以便进行预测和推断。

4.1.1 回归模型

回归模型是通过回归分析建立的数学模型，用来描述自变量与因变量之间的关系。通常情况下，回归模型可以分为一元回归模型和多元回归模型：

- **一元回归模型**：涉及单一自变量与因变量之间的关系。
- **多元回归模型**：涉及多个自变量与因变量之间的关系。

回归模型分类

根据回归模型的特性，可以分为以下几类：

- **线性回归**：自变量与因变量之间呈现线性关系。
- **非线性回归**：自变量与因变量之间呈现非线性关系。
- **多元回归**：包含多个自变量的回归模型。
- **一元回归**：仅包含一个自变量的回归模型。

4.1.2 回归分析的步骤

使用回归分析来解决问题的常规步骤如下：

1. **数据收集**：收集与问题相关的因变量和自变量的数据。
2. **探索性分析**：观察自变量与因变量之间的相关性，并通过图表等方式进行初步分析。
3. **模型构建**：选择合适的回归模型，并根据数据拟合回归方程。
4. **模型评估**：通过统计检验和其他方法，评估回归模型的拟合效果和预测能力。

4.1.3 回归模型的评价指标

回归模型的评价指标用于衡量模型拟合的好坏以及模型的预测能力。常用的指标包括：

- **判定系数（R²）**：表示模型对数据的解释能力，数值越接近1，模型拟合效果越好。

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- **平均绝对误差（MAE）**：对异常值不敏感，直接反映平均误差大小，适用于所有回归模型。

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

- **均方根误差（RMSE）**：对较大误差敏感（因平方放大异常值），与数据规模相关，适用于横向比较相同数据集的模型

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

- **平均绝对百分比误差（MAPE）**：提供百分比误差，便于不同规模数据的比较，对接近零的实际值敏感（可能导致无穷大误差）

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\%$$

指标	公式	范围	异常值敏感度	适用模型	优点	缺点
R²	$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$	[0, 1]	低	线性回归	直观反映解释变量贡献	不适用于非线性模型

指标	公式	范围	异常值敏感度	适用模型	优点	缺点
MAE	$\frac{1}{N}\sum\|\hat{y}_i - y_i\|$	[0, +∞)	低	所有回归模型	鲁棒性强，易解释	忽略误差方向，对大误差不敏感
RMSE	$\sqrt{\frac{1}{N}\sum(\hat{y}_i - y_i)^2}$	[0, +∞)	高	所有回归模型	强调大误差，与标准差单位一致	对异常值敏感，需同数据规模比较
MAPE	$\frac{1}{N}\sum\left\|\frac{\hat{y}_i - y_i}{y_i}\right\| \times 100\%$	[0, +∞)	高（若y≈0）	所有回归模型	无量纲，跨数据可比性	零值无效，小实际值误差放大

4.2 线性回归

线性回归（Linear Regression）

线性回归用于描述**一个因变量与多个自变量**之间的关系。

回归模型采用线性表达式： $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$ 。其中，β_0 是截距项，$\beta_1, \beta_2, \ldots, \beta_p$ 是回归系数，ϵ 是误差项。

4.2.1 一元线性回归（Simple Linear Regression）

简单线性回归模型表示因变量与一个自变量之间的线性关系，公式如下： $y = \beta_0 + \beta_1 x + \epsilon$ ，其假设误差项服从正态分布，且期望值为零，方差为常数。

回归模型的假设

- **线性关系**：自变量与因变量之间的关系为线性。
- **误差项假设**：误差项服从正态分布，且均值为零，方差为常数。
- **独立性假设**：误差项相互独立，且与自变量无关。

线性回归作为最基础的回归分析方法，尽管简单，却能有效地解决许多实际问题。它的数学简洁性使得其在各个学科中得到了广泛应用。尤其是在数据科学中，作为一种预测和分析工具，它有着不可替代的地位。然而，线性回归的假设条件在很多实际数据中难以完全满足。例如，误差项的正态性和独立性往往难以确保，这时可能需要对模型进行调整，如引入正则化方法或使用其他更复杂的回归技术。

一元线性回归的参数估计

普通最小二乘法（OLS）通过构建误差平方和最小化的目标函数，寻找数据与线性模型的最佳匹配路径。其数学本质是在高维空间中寻找最佳的超平面（二维时为直线），使得所有观测点到该平面的垂直距离平方和最小。

为了估计模型中的回归参数（β_0 和 β_1），我们使用最小二乘法（OLS），该方法的核心是最小化残差平方和（RSS）：

$$RSS = \sum_{i=1}^{N}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

其中，y_i 是观测值，$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是回归系数的估计值，x_i 是自变量的观测值。

通过对RSS进行偏导数求解，我们可以得到回归系数的估计值。

参数估计的几何意义

参数估计公式 $\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$ 揭示了变量关系的动态平衡机制：

- 分子项体现协方差特征：数据点分布模式决定回归斜率的符号和强度
- 分母项反映自变量的变异程度：解释变量的离散程度直接影响参数估计的稳定性

线性回归模型的统计假设检验

F检验是用来比较回归模型和无效模型（即仅含常数项的模型）之间的拟合优度差异。其统计量F由以下公式计算：

其中，ESS表示回归平方和（Explained Sum of Squares），RSS表示残差平方和（Residual Sum of Squares），N为样本量，1是回归模型中参数的数量（包括截距项）。

回归模型的统计检验主要包括 **F检验** 和 **t检验**，用于评估回归方程及其系数的显著性。

- **F检验**：$F = \frac{ESS/1}{RSS/(N-2)}$检验整个回归模型是否显著，即解释变量是否整体上对因变量有影响,F值用于评整个回归模型是否显著，比较计算出的F值和临界值 $F_{(1,N-2)}$，若 $F > F_{(1,N-2)}$，则拒绝原假设，认为回归方程显著。
- **t检验**：$t = \frac{\beta}{SE(\beta)}$检验单个回归系数是否显著，判断某个变量是否对因变量有影响，t值用于评估单个回归系数是否显著，比较计算出的 t 值和临界 t 值 $t_{(N-2)}$，若 $|t| > t_{(N-2)}$，则拒绝原假设，认为回归系数显著。

显著性检验通常基于设定的 **显著性水平（α）**，常见取值为 0.05 或 0.01。计算 **p值** 并与 α 比较，p 值越小，回归模型或回归系数越可能是显著的。

比较与思考

- **F检验 vs. t检验**：F检验关注整体模型，t检验关注单个变量。F检验通过时，并不意味着所有变量都是显著的，还需进一步进行 t检验。
- **p值的意义**：p值并不是置信度，而是"在原假设成立的前提下，观察到当前数据或更极端数据的概率"。p值低表示原假设不太可能成立，所以我们倾向于拒绝它。
    - **显著性 ≠ 重要性**：某个变量可能在统计上显著，但实际影响很小。
    - **避免过度依赖显著性检验**：检验结果受样本量影响，样本过大可能导致不重要的变量也显著，需结合实际业务背景分析。

4.2.2 一元线性回归应用案例

本案例通过使用直播数据集（`livestreaming.csv`），进行一元线性回归分析，探讨了 **观看人数（viewers_count）** 和 **主播数量（streamers_count）** 之间的关系。目标是建立一个线性模型，利用主播数量来预测观看人数。

数据与模型构建

- **数据读取与处理**：
    - 使用 `pandas` 读取数据集。
    - 提取出需要的自变量（`streamers_count`）和因变量（`viewers_count`）。

- **添加常数项**：
  - 使用 `sm.add_constant()` 方法为回归模型添加截距项。
- **回归模型创建**：
  - 使用 `sm.OLS()` 创建 **最小二乘法（OLS）** 模型，并通过 `fit()` 方法进行拟合。

结果分析

从模型输出的 **OLS Regression Results** 中得到以下关键信息：

回归模型的系数为：

- 截距项（const）：2.448×10^7
- 主播数量（streamers_count）：9221.6550

因此，回归方程为：

$y = 2.448 \times 10^7 + 9221.6550X$

这里的 X 代表主播数量，y 代表观看人数。

显著性检验

- **p值**：$p = 0.000$（远小于 0.05），表明回归模型中的回归系数在统计上显著，拒绝原假设（即认为系数不为0）。
- **R²**：$R^2 = 0.896$，说明模型解释了大约 89.6% 的观看人数变化。
- **F统计量**：$F = 2248$，表示回归模型整体显著。
- **Durbin-Watson值**：0.260，表明存在一定的自相关问题。
- **Jarque-Bera检验**：$p = 0.0324$，表明残差不完全符合正态分布。

回归模型的拟合效果通过散点图和回归直线展示。图中显示了主播数量与观看人数之间的线性关系，模型拟合较好,该回归模型能够有效预测观看人数，且从统计结果来看，主播数量是观看人数的一个显著影响因素。

4.2.3 多元线性回归

多元线性回归概述

回归模型的数学表达式

在多元线性回归模型中，因变量 y 与多个自变量 $x_1, x_2, \ldots, x_p$ 之间的关系可以表示为：

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$

多元回归模型假设因变量的变化可以通过多个自变量的变化来解释。这些自变量可能相互之间有联系，因此，在建模时需要考虑这些因素对因变量的共同作用。例如，如果我们用浏览历史、点击行为等变量来预测消费者的购买行为，这些变量可能不仅独立地影响购买行为，还可能存在相互作用。

- **因变量（y）**：试图预测或解释的变量。
- **自变量（$x_1, x_2, \ldots, x_p$）**：是对因变量有影响的多个因素。每个自变量的回归系数表示它对因变量的影响大小。
- **回归系数（β）**：每个自变量都有一个对应的回归系数，它衡量自变量对因变量的影响程度。

- **误差项（ϵ）**：表示因变量的部分变化是由无法预见的因素引起的，即模型没有解释的部分。

多元线性回归的关键在于如何通过多个因素预测一个结果。在实际应用中，这种方法常用于解决更复杂的预测问题，比如市场营销、金融风险评估等，因为现实世界中的变量通常并非孤立存在，而是彼此交织的。因此，理解各个自变量如何共同影响因变量，以及它们之间的关系，成为建模和分析过程中的重要步骤。

参数估计

多元线性回归模型的核心在于通过回归系数来建模自变量与因变量之间的关系。假设有 N 个观测数据，每个观测数据包含 p 个自变量，模型的基本形式是：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

将回归模型转换为矩阵形式，可以更方便地进行参数估计。假设有 N 个观测数据，可以表示为以下矩阵形式：

$$Y = X\beta + \epsilon Y = X\beta + \epsilon$$

最小二乘法求解回归系数

多元线性回归的目标是通过**最小化残差平方和（RSS）**来估计回归系数。残差平方和可以表示为：

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

通过最小化 RSS，可以求解回归系数。对 RSS 关于回归系数 β 求导并令其为零，可以得到正态方程：

$$X^T X^\beta = X^T Y$$

解该方程得到回归系数的估计值：

$$\beta^{=} (X^T X)^{-1} X^T Y$$

- **回归系数的估计**：通过最小二乘法，我们得到回归系数的估计值，这些系数表明了自变量对因变量的影响大小。
- **矩阵方法的优势**：通过矩阵表示，可以将多元回归问题转化为线性代数问题，利用矩阵运算可以快速高效地求解回归系数。
- **最小二乘法的原理**：最小二乘法通过最小化误差的平方和来找到最优的回归系数，确保模型的拟合效果最优。

多元线性回归模型的统计检验

检验方法	目的	公式	假设	统计量	结论
拟合优度检验	检验模型拟合效果，表示模型解释因变量变化的程度	$R^2 = 1 - \frac{RSS}{TSS}$	无	R^2	R^2接近1时，模型拟合良好；接近0时，模型拟合差。
F检验	检验回归模型整体显著性	$F = \frac{ESS/p}{RSS/(N-p-1)}$	$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$	$F \sim F(p, N-p-1)$	如果 F 值大于临界值，拒绝 H_0，表示回归模型显著；否则，认为模型不显著。
t检验	检验每个回归系数的显著性	$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$	$H_0 : \beta_j = 0$	$t_j \sim t(N-p-1)$	如果如果$\|t_j\|$大于临界值，则认为回归系数β_j显著，说明自变量x_j对因变量有影响。

4.2.4 多元线性回归案例分析

案例背景

本案例使用 **Auto MPG 数据集**，目标是通过多个自变量（如发动机排量、马力、重量等）预测汽车的燃油效率（`mpg`）。数据中有398条记录，包含多个变量。

数据处理与可视化

数据经过清理后，绘制了自变量之间的相关性热图，发现：

- `displacement` 和 `horsepower` 之间有很强的正相关（0.95）。
- `mpg` 与 `weight` 负相关（-0.78）。

模型建立

使用 `statsmodels` 中的 `OLS()` 方法进行回归建模，将 `mpg` 作为因变量，其他变量作为自变量。

结果分析

- **R^2值**：模型的 $R^2 = 0.82$，说明模型能解释82%的变化。
- **F检验**：$F = 2.73 \times 10^{-14}$，表示模型整体显著，p值小于 0.05，拒绝 H_0。
- **回归系数的显著性**：所有回归系数的 p 值均小于 0.05，表明每个变量对 `mpg` 有显著影响。

回归方程：

$$y = -16.6939 + 0.0114x_1 - 0.0219x_3 - 0.0063x_4 + 0.7484x_6 + 1.3853x_7$$

其中：

- `displacement`（排量）对燃油效率有正向影响。
- `horsepower`（马力）和 `weight`（汽车重量）对燃油效率有负向影响。
- `model_year` 和 `origin` 也对燃油效率有显著影响。

模型能够有效预测燃油效率，解释了大部分的变化。通过回归系数分析，我们可以发现汽车的重量和马力对燃油效率的负面影响，而排量则有正向影响。该模型可用于汽车设计优化和购买决策分析。

4.3 多重共线性

4.3.1 多重共线性概述

多重共线性（Multicollinearity）指的是回归模型中自变量之间存在高度线性相关的现象。具体来说，当一个自变量可以通过其他自变量的线性组合来解释时，就出现了共线性。简单地说，就是在多元回归模型中，自变量之间存在线性依赖关系，导致模型的回归系数无法被准确估计。

例如，如果回归模型中的自变量 $x_1, x_2, \ldots, x_p$ 存在共线性，可以用以下关系表示：

$$c_0 + c_1x_1 + c_2x_2 + \cdots + c_px_p = 0$$

如果系数 $c_0, c_1, \ldots, c_p$ 不为零，则说明存在完全多重共线性。

4.3.2 识别多重共线性

方法	目的	公式	适用场景	优点	缺点
相关矩阵（Correlation Matrix）	检查自变量之间的相关性	计算各自变量之间的皮尔逊相关系数	检查自变量之间的线性关系，发现高度相关的变量	简单易用，直观地显示变量间的相关性	仅能揭示明显的相关性，无法提供强度或程度的信息
方差膨胀因子（VIF）	衡量自变量之间的共线性程度	$VIFj = \frac{1}{1-R_j^2}$	检查每个自变量与其他自变量的线性关系程度	提供自变量的共线性程度，能够识别问题变量	对于高度共线性的情况，VIF值可能非常大，容易误解或忽略潜在问题
条件数（Condition Number）	衡量设计矩阵的条件数，检测自变量间的多重共线性	通过矩阵的特征值分解计算最大特征值与最小特征值之比	检查设计矩阵是否存在共线性	能够揭示共线性引起的数值不稳定性，适合处理高维数据	不易于理解，且只适用于高维数据，低维数据可能无法有效应用
主成分分析（PCA）	将自变量降维，将相关性较高的自变量合并成主成分	通过特征值分解计算主成分	自变量间存在高共线性时，采用主成分替代原有变量	可减少多重共线性，避免冗余自变量问题	可能导致解释性降低，因为主成分并不总是与实际变量直接相关
偏最小二乘法（PLS）	用于高共线性和小样本的回归分析，减少共线性影响	构建主成分，通过最小二乘法回归分析	高共线性且样本量较少时，作为回归建模的选择	能处理多重共线性问题，同时提高模型的稳定性	在解释模型时，可能不如标准回归模型直观和易理解

方法	目的	公式	适用场景	优点	缺点
逐步回归（Stepwise Regression）	通过自动选择重要的自变量来降低多重共线性	基于AIC或BIC等准则，选择或剔除自变量	在大规模回归分析中，逐步筛选显著变量	自动化处理，简化变量选择过程	可能过度拟合或忽略一些边缘重要的变量，且结果受初始选择的影响大

4.3.3 多重共线性的解决方案

方法	目的	适用场景	优点	缺点
删除多重共线性变量	删除相关性较强的自变量以减轻共线性影响	自变量间高度相关，需剔除其中一个或多个变量	简单有效，容易实施	可能导致丢失重要的解释变量，影响模型准确性
岭回归（Ridge Regression）	通过增加正则化项来减少回归系数的波动，减轻共线性问题	自变量存在多重共线性时，尤其是模型不稳定时	能有效减小回归系数的方差，减轻多重共线性影响	可能导致模型的解释性下降，特别是当正则化项较大时
主成分分析（PCA）	通过将相关自变量合并为少数几个主成分来减小共线性	高度相关的自变量，减少变量数量并提取主要信息	能有效减少共线性，提高模型稳定性，便于数据降维	解释性较差，转换后的主成分可能难以与原始变量直接关联
增加样本量（N）	通过增加样本量来减小共线性带来的影响	当数据集较小，且样本量不足时	简单直观，能够通过增加数据量提高模型的鲁棒性	难以在现有数据基础上进行，收集更多数据可能成本较高
偏最小二乘法（PLS）	通过生成新的主成分替代原始变量来降低多重共线性影响	高共线性且样本量较少时	可以处理多重共线性问题，模型稳定性较好	解释性差，模型较为复杂，可能导致信息损失

4.4 岭回归

4.4.1 岭回归概述

岭回归（Ridge Regression）是一种带有L2正则化项的回归方法，它通过在损失函数中加入正则化项来减少回归系数的方差，从而解决多重共线性问题。岭回归特别适用于自变量之间高度相关的情况，避免了普通最小二乘法（OLS）在共线性情况下的系数估计不稳定问题。

损失函数：

岭回归的目标是最小化以下损失函数：$\min_\beta \|y - X\beta\|^2 + \lambda\|\beta\|^2$

岭回归求解

通过最小化损失函数，我们可以得到岭回归的解。损失函数的最小化过程如下：
$L = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$

对损失函数对回归系数 β 求导并令其为零，得到如下的方程：

$\frac{\partial \mathcal{L}}{\partial \beta} = -2X^T(Y - X\beta) + 2\lambda\beta = 0$

简化后得到：$(X^TX + \lambda I)\beta = X^TY$

通过解上述方程，可以得到岭回归的回归系数：$\hat{\beta} = (X^TX + \lambda I)^{-1}X^TY$

这个公式表示，岭回归通过调整 X^TX 的逆矩阵，加上正则化项 λI，来得到更稳定的回归系数估计。

4.4.2 岭回归参数 λ 的选择

- **岭迭代法**：这种方法通过绘制不同岭参数下岭回归模型的系数变化曲线，从而通过观察系数的稳定性来确定最佳的 λ 值。通过绘制岭回归模型的系数变化图，可以根据曲线的形状和趋势来判断合适的 λ 值，从而确定目标的 λ。
- **交叉验证**：交叉验证是一种常见的模型选择方法，通过将数据集划分为若干个子集（通常是 k 折交叉验证），然后进行 k 次模型训练和评估。交叉验证可以评估不同岭回归模型在每个子集上的性能，并根据平均性能选择最佳的 λ 值。通常，这种方法使用均方误差（MSE）等评估指标来比较各个候选模型的性能。

4.4.3 岭回归案例研究：基于岭回归的汽车燃油效率预测案例

案例背景与数据处理

本案例使用了汽车数据集来进行燃油效率（`mpg`）的预测。数据集包含多个变量，如 `displacement`（发动机排量）、`horsepower`（马力）、`weight`（重量）、`acceleration`（加速能力）等。数据被分为训练集和测试集，使用 `train_test_split` 函数进行数据的划分，比例为 70% 的训练集和 30% 的测试集。数据经过标准化处理，即将每个特征减去均值并除以标准差。

岭回归建模与评估

在数据预处理后，使用 `Ridge` 类对训练集进行拟合，并计算回归系数。使用了岭回归的 `alpha=1.0` 参数，这意味着正则化项的强度适中。岭回归通过正则化减少了自变量间的多重共线性问题，避免了普通最小二乘法（OLS）可能出现的过拟合问题。

VIF（方差膨胀因子）

在模型构建前，通过计算VIF来评估各自变量之间的共线性问题。VIF值较高的变量表明这些变量可能存在线性关系，因此需要特别注意。通过 `variance_inflation_factor` 函数计算每个自变量的VIF值，得到了以下结果：`displacement`，`cylinders`，`weight` 等变量的VIF值较高，说明它们之间有较强的相关性。

模型评估

通过计算训练集和测试集上的均方根误差（RMSE）和 R^2 值对模型进行了评估。模型预测结果的RMSE为3.43，R^2值为0.824，表明该岭回归模型能够较好地拟合数据，具有较高的解释能力。

岭回归的回归方程为：

$$y = 23.414 + 1.688x_1 - 0.284x_2 - 0.17x_3 - 5.84x_4 + 0.707x_5 + 2.708x_6 + 1.09x_7$$

其中各个自变量（`displacement`，`horsepower`，`weight`，`acceleration`，`model_year`，`origin`）对燃油效率（`mpg`）的贡献通过回归系数体现。

- **发动机排量（displacement）**对燃油效率有正向影响，即排量越大，燃油效率越高。
- **汽车重量（weight）** 和 **马力（horsepower）** 对燃油效率有负向影响，表明较重和马力较大的汽车通常消耗更多的燃料。
- **加速能力（acceleration）**和**生产年份（model_year）**对燃油效率也有一定影响。

岭回归效地处理了自变量间的多重共线性问题，并成功预测了汽车的燃油效率。通过正则化，模型避免了过拟合，提供了稳定且合理的回归系数。最终的模型评估结果显示，岭回归具有较高的预测能力，并且能够有效识别各个变量对燃油效率的影响。这一案例展示了如何应用岭回归处理多重共线性问题，并通过模型评估来验证其效果。

4.5 LASSO回归

4.5.1 LASSO回归概述

核心概念

- **定义**：LASSO（Least Absolute Shrinkage and Selection Operator）回归是一种通过L1正则化最小化残差平方和的线性回归方法，其目标函数为：
  $\min_\beta \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$
  其中 $\lambda > 0$ 控制正则化强度。
- **与岭回归的差异**：
  - 岭回归（L2正则化）使系数接近但不为0，无法筛选变量。
  - LASSO（L1正则化）可将不重要变量的系数压缩至0，实现**变量筛选**，适用于高维数据和共线性场景。

4.5.2 坐标下降法求解

方法原理

- **核心思想**：逐分量更新系数，每次仅优化一个变量，其余固定。
- **残差平方和（RSS）**：
  $RSS(\beta) = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$
  展开后对 β_j 求偏导，结合L1正则化的次梯度条件推导更新规则。

数学推导

1. **中间变量定义**：
   - $m_j = \sum_{i=1}^{N} x_{ij} \left(y_i - \sum_{k \neq j} \beta_k x_{ik} \right)$（调整后的协方差项）
   - $n_j = \sum_{i=1}^{N} x_{ij}^2$（特征 j 的平方和）
2. **次梯度条件**：
   L1正则化项的次导数为：
   $$\frac{\partial \lambda |\beta_j|}{\partial \beta_j} = \begin{cases} \lambda & \beta_j > 0 \\ [-\lambda, \lambda] & \beta_j = 0 \\ -\lambda & \beta_j < 0 \end{cases}$$
   结合RSS的偏导，得到系数更新条件：
   $$\beta_j = \begin{cases} \frac{m_j - \lambda/2}{n_j}, & m_j > \frac{\lambda}{2} \\ 0, & |m_j| \leq \frac{\lambda}{2} \end{cases}$$

算法流程

1. 初始化系数 β。
2. 循环遍历每个特征 $j = 1, 2, \ldots, p$，按条件更新 β_j。
3. 迭代至收敛或达到最大步数。

参数选择

- λ**的作用**：
  - λ 越大，更多系数被压缩至0，模型更稀疏。
  - 最优 λ 通过交叉验证确定（如K折交叉验证）。

4.5.3 LASSO回归实践案例：汽车燃油效率预测

数据处理与模型构建

基于4.2节的数据进行LASSO回归分析，数据预处理步骤与4.4.3节相同。使用`sklearn.linear_model`中的`Lasso`类来构建回归模型，主要参数`alpha`表示LASSO中的正则化系数。

结果评估

LASSO回归结果包括系数、截距、R^2 值和均方根误差 (RMSE)。通过绘制散点图和残差图来评估模型效果。

回归结果

最终回归模型为：

$$y = 23.414 - 4.86x_4 + 0.47x_5 + 2.58x_6 + 0.82x_7$$

其中，`x_1` 为汽车重量，`x_2` 为引擎气缸数，`x_3` 为发动机马力，`x_4` 为车辆年份。

模型评估

- **RMSE** = 3.4228
- R^2 = 0.824

这表明LASSO回归模型很好地拟合了数据，能够有效预测燃油效率。LASSO回归通过L1正则化选择了最重要的特征，有助于减少过拟合。

案例启示

通过与传统多元线性回归模型对比，LASSO回归表现出更好的性能，能够有效处理特征之间的多重共线性问题。其在数据维度较高时尤为有效，能够通过正则化减少模型复杂度。

4.6 非线性回归

非线性回归是一种用于处理自变量和因变量之间非线性关系的回归分析方法。在非线性回归中，自变量和因变量之间的关系可以是曲线、指数、对数、幂函数等非线性形式。非线性回归模型包括本质非线性模型和扩展非线性模型，具体包括：

- **本质非线性模型**（函数混合效应模型）
- **扩展非线性模型**（生长曲线模型、广义线性模型、混合效应模型）

4.6.1 本质非线性模型

本质非线性模型通过对自变量进行非线性变换，旨在将原本的非线性关系转化为线性关系。常见的几种本质非线性模型及其变换方式如下：

4.6.1.1 常见的本质非线性模型

概念名称	原方程式	变换思路	变换后方程式
三次曲线	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$	使用数学变换，$x = z^3$	$y = \beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3$
复合曲线	$y = \beta_0 + \beta_1 e^{\beta_2 x}$	对数变换	$\ln(y - \beta_0) = \beta_1 + \beta_2 x$
增长曲线	$y = \beta_0 e^{-\beta_1 x}$	使用数学变换，$x = z^{-1}$	$\ln(y) = \ln(\beta_0) - \beta_1 z$

其他常见变换模型

概念名称	原方程式	变换思路	变换后方程式
对数曲线	$y = \beta_0 + \beta_1 \ln(x)$	使用数学变换，$x = \ln(z)$	$y = \beta_0 + \beta_1 z$
三次曲线	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$	使用数学变换，$x = \ln(z)$	$\ln(y) = \ln(\beta_0) + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

概念名称	原方程式	变换思路	变换后方程式
指数曲线	$y = \beta_0 e^{\beta_1 x}$	对数变换	$\ln(y) = \ln(\beta_0) + \beta_1 x$

非线性回归在很多实际问题中非常有用，尤其是当自变量与因变量之间的关系无法用线性回归表达时。通过适当的变换，可以将非线性问题转化为线性问题，从而简化模型的计算和分析过程。

4.6.1.2 多项式回归

多项式回归通过引入自变量的多项式项来拟合因变量与自变量之间的非线性关系。一般的回归模型形式如下：

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m + \epsilon$$

回归类型	公式表达式	应用场景
一元多项式回归	$y = \beta_0 + \beta_1 x + \cdots + \beta_m x^m + \epsilon$	用于拟合单一自变量与因变量之间的关系，常用于趋势预测
多元多项式回归	$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$	用于多自变量的情况，常用于经济学、工程等领域的建模

4.6.2 本质非线性模型

本质非线性模型指的是通过变换自变量，使得回归问题能够以线性模型进行处理。其数学表达式如下：

$$y = g(x, \beta) + \epsilon$$

其中，$g(x, \beta)$ 是自变量x与参数β之间的非线性函数，$i = 1, 2, \ldots, N$，$N > p$，p为自变量的维度。

为了估计回归系数β，我们使用最小二乘法进行拟合。目标是最小化以下损失函数：

$$\text{argmin}_\beta \left(Q(\beta) = \sum_{i=1}^{N} \left(y_i - g(x_i, \beta) \right)^2 \right)$$

通过对损失函数求导，得到以下方程：

$$\partial Q(\beta) \partial \beta j = -2 \sum_{i=1}^{N} \left(y_i - g(x_i, \beta) \right) \frac{\partial g(x_i, \beta)}{\partial \beta_j} = 0 \quad j = 1, 2, \ldots, p$$

通过Newton优化法求解该方程，最终得到最优的β

4.6.3 基于工龄的月薪预测非线性回归案例笔记

案例背景

- **数据来源**：IT行业100名员工的工龄（年）与月薪（元）数据，工龄范围0-8年，月薪范围约10,808至36,053元。
- **目标**：探究工龄与月薪的非线性关系，比较二次曲线与幂函数模型的拟合效果。

模型方法:二次曲线模型($y = \beta_0 + \beta_1 x + \beta_2 x^2$)和幂函数模型（($y = a \cdot x^b$)

结果对比与分析

指标	二次曲线模型	幂函数模型
实现方式	使用`PolynomialFeatures(degree=2)`生成二次特征，通过线性回归（`LinearRegression`）拟合扩展后的多项式特征。	自定义幂函数`power_func(x, a, b)`，利用`curve_fit`优化参数。
结果模型	$y = 400.80x^2 - 743.68x$	$y = 12345.70 \cdot x^{0.4}$
R^2	0.931	0.706
拟合特点	捕捉非线性增长趋势	描述平缓增长关系

关键结论

- **二次曲线优势**：
  - R^2值显著更高（93.1%），拟合曲线更贴合数据点。
  - 工龄对月薪的影响呈现**加速增长**趋势（二次项系数为正），符合实际中经验积累带来的薪资跃升规律。
- **幂函数局限性**：
  - 模型假设为单调增长，但指数$b = 0.399$反映增速放缓，无法捕捉后期薪资的快速上升。
- **业务意义**：
  - 工龄与月薪的关系具有明显非线性特征，需优先选择能描述复杂变化的模型（如多项式回归）。
  - 二次模型可为HR薪资结构设计提供量化依据，例如预测高工龄员工的薪资涨幅。

可视化与验证

- **散点图与拟合曲线**：
  - 二次曲线在工龄较高区域（>5年）预测值显著上升，与数据分布一致。
  - 幂函数曲线整体平缓，低估高工龄员工的薪资水平。
- **模型诊断**：
  - 二次模型残差较小且分布均匀，幂函数残差随工龄增加而扩大。

案例启示

- **模型选择**：非线性问题需优先尝试多项式回归或更灵活的模型（如样条回归）。
- **可解释性**：二次模型明确量化了工龄的边际效应（一阶项负向、二阶项正向），便于业务解读。
- **扩展方向**：可引入更多特征（如岗位类别、技能等级）构建多元非线性模型。